

# AR Identification of Latent-variable Graphical Models

Mattia Zorzi, Rodolphe Sepulchre

## Abstract

The paper proposes an identification procedure for autoregressive gaussian stationary stochastic processes wherein the manifest (or observed) variables are mostly related through a limited number of latent (or hidden) variables. The method exploits the sparse plus low-rank decomposition of the inverse of the manifest spectral density and the efficient convex relaxations recently proposed for such decomposition.

## Index Terms

Latent-variable graphical models, system identification, convex relaxation, convex optimization.

## I. INTRODUCTION

Gaussian processes and their representation by graphical models have gained popularity through science and engineering, [1], [2]. The objective of the present paper is to derive an identification procedure for gaussian stochastic processes whose manifest (observed) variables are correlated primarily through a restricted number of latent (hidden) variables. Here, (the few) latent variables are fictitious elements introduced by the modeler. The resulting graphical model (or equivalently

This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its authors. This research is also supported by FNRS (Belgian Fund for Scientific Research).

M. Zorzi is with the Department of Electrical Engineering and Computer Science, University of Liège, 4000 Liège, BE, (mzorzi@ulg.ac.be)

R. Sepulchre is with the Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, UK, (r.sepulchre@eng.cam.ac.uk), and the Department of Electrical Engineering and Computer Science, University of Liège, 4000 Liège, BE.

latent-variable graphical model) has a two layer structure, one layer for the manifest (observed) nodes and one layer for the latent (hidden) nodes. The hope is that in many applications of interest, the few extra nodes in the hidden layer allow for a drastic reduction of edges in the observed layer, because the observed nodes become nearly independent when conditioned to the hidden nodes. As a consequence, allowing for latent variables in the identification of the stochastic model may improve scalability and robustness of the algorithm. This paradigm was exploited in the framework of gaussian random vectors in the recent paper [3]. The authors exploited the sparse plus low-rank (S+L) decomposition of the manifest concentration matrix (the inverse of the covariance matrix corresponding to the manifest variables) to propose an efficient formulation of the identification problem.

The present paper focuses on the generalization of this approach to autoregressive (AR) gaussian stationary processes, exploiting the analog sparse plus low-rank decomposition of the inverse of the manifest spectral density (the spectral density of the manifest variables). It thereby connects the extensive recent research on convex regularization of sparsity and low-rank constraints [3], [4], [5], [6], [7] to the classical covariance extension approach for the identification of gaussian stationary processes [8], [9]. It also provides a generalization of recent contributions that introduced sparsity constraints (but no latent variables) in the identification of autoregressive processes [10], [11], [12].

The paper is organized as follows. After mathematical preliminaries, Section II introduces the main ideas of the proposed identification scheme in non technical terms. The identification of the graphical model and the identification of the autoregressive model are formulated as two distinct optimization problems. The first one uses sparsity and low-rank regularizers to recover the model structure. It is further analyzed in Section III. The second one solves an exact covariance extension problem for a fixed graphical model. It is further analyzed in Section IV. Finally, in Section V we discuss an illustrative example and test our method to international stock return data.

### *Notation*

We endow the vector space  $\mathbb{R}^{m \times m}$  with the usual inner product  $\langle X, L \rangle = \text{tr}(XL^T)$ .  $\mathbf{Q}_m$  denotes the vector space of symmetric matrices of dimension  $m$ , if  $X \in \mathbf{Q}_m$  is positive definite (semi-definite) we write  $X \succ 0$  ( $X \succeq 0$ ). A matrix  $A \in \mathbb{R}^{l \times m(n+1)}$  with  $l \leq m$  will be

partitioned as  $A = \begin{bmatrix} A_0 & A_1 & \dots & A_n \end{bmatrix}$  with  $A_j \in \mathbb{R}^{l \times m}$ .  $\mathbf{M}_{m,n}$  is the vector space of matrices  $Y := \begin{bmatrix} Y_0 & Y_1 & \dots & Y_n \end{bmatrix}$  with  $Y_0 \in \mathbf{Q}_m$  and  $Y_1 \dots Y_n \in \mathbb{R}^{m \times m}$ . The corresponding inner product is  $\langle Y, Z \rangle = \text{tr}(Y Z^T)$ . The linear mapping  $T : \mathbf{M}_{m,n} \rightarrow \mathbf{Q}_{m(n+1)}$  constructs a symmetric *Toeplitz* matrix from its first block row in the following way:

$$T(Y) = \begin{bmatrix} Y_0 & Y_1 & \dots & Y_n \\ Y_1^T & Y_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & Y_1 \\ Y_n^T & \dots & Y_1^T & Y_0 \end{bmatrix}. \quad (1)$$

The adjoint of  $T$  is a mapping  $D : \mathbf{Q}_{m(n+1)} \rightarrow \mathbf{M}_{m,n}$  defined as follows. If  $X \in \mathbf{Q}_{m(n+1)}$  is partitioned as

$$X = \begin{bmatrix} X_{00} & X_{01} & \dots & X_{0n} \\ X_{01}^T & X_{11} & \dots & X_{1n} \\ \vdots & \vdots & & \vdots \\ X_{0n}^T & X_{1n}^T & \dots & X_{nn} \end{bmatrix} \quad (2)$$

then  $D(X) = \begin{bmatrix} D_0(X) & \dots & D_n(X) \end{bmatrix}$  where

$$D_0(X) = \sum_{h=0}^n X_{hh}, \quad D_j(X) = 2 \sum_{h=0}^{n-j} X_{h \ h+j}, \quad j = 1 \dots n. \quad (3)$$

We define the index set  $E_m \subseteq V_m \times V_m$  with  $V_m := \{1, 2, \dots, m\}$ , and its complement set is denoted by  $E_m^c$ . The cardinality of  $E_m$  is denoted by  $|E_m|$ . The projection map  $P_{E_m} : \mathbb{R}^{m \times m} \rightarrow \mathbb{R}^{m \times m}$  is defined as follows

$$P_{E_m}(X) = \begin{cases} (X)_{kh}, & (k, h) \in E_m \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where  $(X)_{kh}$  is the entry of  $X$  in position  $(k, h)$ . Similarly,  $P_{E_m}(Y)$  with  $Y \in \mathbf{M}_{m,n}$  denotes

$$\begin{bmatrix} P_{E_m}(Y_0) & P_{E_m}(Y_1) & \dots & P_{E_m}(Y_n) \end{bmatrix}. \quad (5)$$

Functions on the unit circle  $\{e^{i\vartheta} \text{ s.t. } \vartheta \in [-\pi, \pi]\}$  will be denoted by capital Greek letters, e.g.  $\Phi(e^{i\vartheta})$  with  $\vartheta \in [-\pi, \pi]$ , and the dependence upon  $\vartheta$  will be dropped if not needed, e.g.  $\Phi$  instead of  $\Phi(e^{i\vartheta})$ .  $L_2^{m \times m}$  denotes the space of  $\mathbb{C}^{m \times m}$ -valued functions defined on the unit circle which are square integrable. Given  $\Phi \in L_2^{m \times m}$ , the shorthand notation  $\int \Phi$  denotes the integration of  $\Phi$  taking place on the unit circle with respect to the normalized *Lebesgue* measure. Then, the

inner product in  $L_2^{m \times m}$  is  $\langle \Phi, \Sigma \rangle = \text{tr} \int \Phi \Sigma^*$ . Similarly,  $P_{E_m} : L_2^{m \times m} \rightarrow L_2^{m \times m}$  is defined as in (4) where  $X$  is replaced by  $\Phi(e^{i\vartheta})$ . Moreover,  $\sigma_k(\Phi(e^{i\vartheta}))$  denotes the  $k$ -th largest singular value of  $\Phi(e^{i\vartheta})$  at  $\vartheta$ , i.e.  $\sigma_1(\Phi(e^{i\vartheta})) \geq \sigma_2(\Phi(e^{i\vartheta})) \geq \dots \geq \sigma_m(\Phi(e^{i\vartheta}))$  for each  $\vartheta \in [-\pi, \pi]$ .  $\mathcal{A}_m$  denotes the linear space of  $\mathbb{C}^{m \times m}$ -valued analytic functions on the unit circle. Given  $\Lambda \in \mathcal{A}_m$ , we define the norm

$$\|\Lambda\| = \sup_{\vartheta \in [-\pi, \pi]} \sigma_1(\Lambda(e^{i\vartheta})) \quad (6)$$

and the (normal) rank

$$\text{rank}(\Lambda) := \max_{\vartheta \in [-\pi, \pi]} \text{rank}(\Lambda(e^{i\vartheta})). \quad (7)$$

If  $\Phi(e^{i\vartheta})$  is positive definite (semi-definite) for each  $\vartheta \in [-\pi, \pi]$ , we will write  $\Phi \succ 0$  ( $\Phi \succeq 0$ ).  $\mathcal{S}_m$  denotes the family of functions  $\Phi$  such that  $\Phi = \Phi^*$  and  $c_1 I \preceq \Phi \preceq c_2 I$  for some  $c_1, c_2 > 0$ . We define the following family of matrix pseudo-polynomials

$$\mathcal{Q}_{m,n} = \left\{ \sum_{j=-n}^n e^{-ij\vartheta} R_j \text{ s.t. } R_j = R_{-j}^T \in \mathbb{R}^{m \times m} \right\}. \quad (8)$$

The *shift operator* is defined as

$$\Delta(e^{i\vartheta}) := \begin{bmatrix} I_m & e^{i\vartheta} I_m & \dots & e^{in\vartheta} I_m \end{bmatrix}. \quad (9)$$

Given  $X \in \mathcal{Q}_{m(n+1)}$ , by direct computation we get

$$\begin{aligned} \Delta(e^{i\vartheta}) X \Delta(e^{i\vartheta})^* \\ = D_0(X) + \frac{1}{2} \sum_{j=1}^n e^{-ij\vartheta} D_j(X) + e^{ij\vartheta} D_j(X)^T, \end{aligned} \quad (10)$$

therefore  $\Delta X \Delta^* \in \mathcal{Q}_{m,n}$ . On the other hand, any element in  $\mathcal{Q}_{m,n}$  may be parameterized as (10) because  $D$  is a surjective map. We conclude that

$$\mathcal{Q}_{m,n} = \{ \Delta X \Delta^* \text{ s.t. } X \in \mathcal{Q}_{m(n+1)} \}. \quad (11)$$

## II. PROBLEM FORMULATION

### A. AR Model Identification

Let  $L_2^m(\Omega, \mathcal{A}, P)$  be the Hilbert space of second order  $\mathbb{R}^m$ -valued gaussian random vectors defined in the probability space  $\{\Omega, \mathcal{A}, P\}$ . An  $\mathbb{R}^m$ -valued gaussian stochastic process  $x^m$  is an ordered collection of random vectors  $x^m = \{x^m(t); t \in \mathbb{Z}\}$  in  $L_2^m(\Omega, \mathcal{A}, P)$ . Moreover, we

assume  $x^m$  is zero mean, stationary and purely nondeterministic. It is completely described by its spectral density

$$\Phi_m(e^{i\vartheta}) = \sum_{j=-\infty}^{\infty} e^{-ij\vartheta} R_j \quad (12)$$

where  $R_j := \mathbb{E}[x^m(t+j)x^m(t)^T]$  denotes the  $j$ -th covariance lag. An empirical estimate  $\hat{R}_j$  of  $R_j$  is computed from a finite-length realization of  $x^m$ , i.e.  $x^m(1), x^m(2), \dots, x^m(N)$ , as follows

$$\hat{R}_j = \frac{1}{N} \sum_{t=0}^{N-j} x^m(t+j)x^m(t)^T. \quad (13)$$

The estimate  $\hat{\Phi}_m^\circ$  of  $\Phi_m$  that maximizes the *entropy rate*, [13], and that matches the first  $n$  covariance lags is the solution of the following convex program [8]:

$$\begin{aligned} \hat{\Phi}_m^\circ &= \arg \max_{\Phi_m \in \mathcal{S}_m} \int \log \det \Phi_m \\ \text{subject to} \quad &\int \Delta \Phi_m = \hat{R} \end{aligned} \quad (14)$$

The matrix  $\hat{R} := \begin{bmatrix} \hat{R}_0 & \hat{R}_1 & \dots & \hat{R}_n \end{bmatrix} \in \mathbf{M}_{m,n}$  satisfies  $\mathbf{T}(\hat{R}) \succ 0$ , [14].  $\hat{\Phi}_m^\circ$  is usually referred to as *maximum-entropy covariance extension*. Because  $(\hat{\Phi}_m^\circ)^{-1} \succ 0$  belongs to  $\mathcal{Q}_{m,n}$ , it admits the spectral factorization  $\hat{\Phi}_m^\circ = \Gamma \Gamma^*$  where  $\Gamma = (A \Delta^*)^{-1}$ ,  $A \in \mathbb{R}^{m \times m(n+1)}$ , is a shaping filter for the estimated process,  $\hat{x}_m^\circ$ , [15]. This means that  $\hat{x}_m^\circ$  is the output of  $\Gamma$  fed by white gaussian noise (WGN), say  $e$ , with zero mean and variance equal to the identity:

$$\hat{x}_m^\circ(t) = \sum_{j=0}^n A_j \hat{x}_m^\circ(t-j) + e(t) \quad (15)$$

therefore the maximum entropy estimate is an autoregressive process. In [16], [17] it has been shown that the dual of (14) is

$$\begin{aligned} \min_{\Phi_m^{-1} \in \mathcal{Q}_{m,n}} \quad &\int \left( -\log \det \Phi_m^{-1} + \left\langle \Phi_m^{-1}, \hat{\Phi}_m \right\rangle \right) \\ \text{subject to} \quad &\Phi_m \succ 0 \end{aligned} \quad (16)$$

where  $\hat{\Phi}_m(e^{i\vartheta}) := \sum_{j=-n}^n e^{-ij\vartheta} \hat{R}_j$  is the  $n$ -length windowed *correlogram* of  $x^m$ , [14]. Note that  $\hat{\Phi}_m$  is not necessarily positive semi-definite on the unit circle.

### B. Spectral Density of Latent-variable Graphical Models

We consider a real, zero-mean, stationary, purely nondeterministic, gaussian process  $x = \{x(t); t \in \mathbb{Z}\}$  with  $m$  manifest variables and  $l$  latent variables, that is  $x := \begin{bmatrix} (x^m)^T & (x^l)^T \end{bmatrix}^T$  where  $x^m := \begin{bmatrix} x_1 & \dots & x_m \end{bmatrix}^T$  and  $x^l := \begin{bmatrix} x_{m+1} & \dots & x_{m+l} \end{bmatrix}^T$ . Let  $I \subset V_{m+l}$  be an arbitrary index set. We denote as

$$\mathcal{X}_I = \overline{\text{span}}\{x_j(t) \text{ s.t. } j \in I, t \in \mathbb{Z}\} \quad (17)$$

the closure in  $L_2^{m+l}(\Omega, \mathcal{A}, P)$  of the vector space of all finite linear combinations (with real coefficients) of  $x_j(t)$  with  $j \in I$  and  $t \in \mathbb{Z}$ , [18, page 3]. The shorthand notation

$$\mathcal{X}_{\{k\}} \perp \mathcal{X}_{\{h\}} \mid \mathcal{X}_{V_{m+l} \setminus \{k,h\}} \quad (18)$$

means that  $\mathcal{X}_{\{k\}}$  and  $\mathcal{X}_{\{h\}}$  are conditionally independent given  $\mathcal{X}_{V_{m+l} \setminus \{k,h\}}$ , see [12]. Therefore, (18) signifies that  $x_k$  and  $x_h$  are conditional independent given the space linearly generated by  $x_j$  with  $j \in V_{m+l} \setminus \{k,h\}$ . Conditional dependence relations among the variables of the process  $x$  define an *interaction graph*  $\mathcal{G} = (V_{m+l}, E_{m+l})$  whose nodes represent the variables  $x_1, x_2, \dots, x_{m+l}$  and edges represent conditional dependence:

$$(k, h) \notin E_{m+l} \iff k \neq h, \mathcal{X}_{\{k\}} \perp \mathcal{X}_{\{h\}} \mid \mathcal{X}_{V_{m+l} \setminus \{k,h\}}. \quad (19)$$

The graph  $\mathcal{G}$  leads to a *latent-variable graphical model* of the gaussian process. It admits the two layer structure illustrated in Figure 1: latent nodes are in the upper level, and manifest nodes

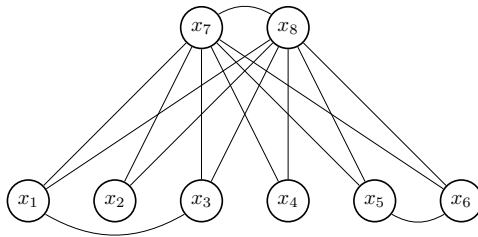


Fig. 1. Example of a latent-variable graphical model:  $x_1, x_2, \dots, x_6$  are manifest variables  $x_7, x_8$  are latent variables.

are in the lower level.

The graphical structure of  $x$  translates into a particular decomposition of its spectral density  $\Phi \in \mathcal{S}_{m+l}$ . Starting from the block decomposition

$$\Phi = \begin{bmatrix} \Phi_m & \Phi_{lm}^* \\ \Phi_{lm} & \Phi_l \end{bmatrix}, \quad \Phi^{-1} = \begin{bmatrix} \Upsilon_m & \Upsilon_{lm}^* \\ \Upsilon_{lm} & \Upsilon_l \end{bmatrix} \quad (20)$$

we obtain the relationship

$$\Phi_m^{-1} = \Upsilon_m - \Upsilon_{lm}^* \Upsilon_l^{-1} \Upsilon_{lm}. \quad (21)$$

where we used the *Schur complement* pointwise.

Our main modeling assumption are that  $l \leq m$  and the conditional dependence relations among the manifest variables are mostly through this limited number of latent variables. This means that the corresponding graphical model  $\mathcal{G}$  has few edges between the manifest nodes, and few latent nodes. This leads to a S+L structure for (21), that is,

$$\Phi_m^{-1} = \Sigma - \Lambda, \quad \Lambda \succeq 0 \quad (22)$$

where  $\Sigma \in \mathcal{Q}_{m,n}$  is sparse and  $\Lambda \in \mathcal{Q}_{m,n}$  is low-rank. This means that the support of  $\Sigma$ , denoted by  $E_m$ , contains few elements, and there exists  $G \in \mathbb{R}^{l \times m(n+1)}$  with  $l \ll m$  and full row rank such that  $\Lambda = \Delta G^T G \Delta^*$ . Accordingly,  $\Phi_m^{-1}$  may be decomposed into the following two finite dimensional vector subspaces

$$\begin{aligned} \mathcal{V}_{E_m} &:= \{\Sigma \in \mathcal{Q}_{m,n} \text{ s.t. } P_{E_m^c}(\Sigma) = 0\} \\ \mathcal{V}_G &:= \{\Delta G^T H G \Delta^* \text{ s.t. } H \in \mathbf{Q}_l\}. \end{aligned} \quad (23)$$

The sparsity of  $\Sigma$  reflects the presence of few edges among the manifest nodes of  $\mathcal{G}$  because of the relationship

$$\begin{aligned} (\Phi(e^{i\vartheta})^{-1})_{kh} &= 0, \quad \forall \vartheta \in [-\pi, \pi] \Leftrightarrow \\ \mathcal{X}_{\{k\}} &\perp \mathcal{X}_{\{h\}} \mid \mathcal{X}_{V_{m+l} \setminus \{k,h\}} \end{aligned} \quad (24)$$

which has been shown in [12], see also [19], [20]. The nonzero entries of  $\Sigma$  therefore correspond to the (few) conditional dependence relations among the manifest variables. Accordingly, the more  $\Sigma$  sparse is, the less conditional dependence relations among the manifest variables we have. Since  $l \leq m$ , the rank of  $\Lambda = \Upsilon_{lm}^* \Upsilon_l^{-1} \Upsilon_{lm}$  coincides with  $l$ , that is the number of latent

variables. Accordingly the more low-rank  $\Lambda$  is, the less latent variables we have. It is worth noting that (21) is a dynamical generalization of the static decomposition

$$R_m^{-1} = K_m - K_{lm}^* K_l^{-1} K_{lm} \quad (25)$$

for a zero mean gaussian random vector  $x = \begin{bmatrix} (x^m)^T & (x^l)^T \end{bmatrix} \sim \mathcal{N}(0, R)$  with

$$R = \begin{bmatrix} R_m & R_{lm}^T \\ R_{lm} & R_l \end{bmatrix}, \quad R^{-1} = \begin{bmatrix} K_m & K_{lm}^T \\ K_{lm} & K_l \end{bmatrix}, \quad (26)$$

see [3]. Finally, in the case  $\Sigma$  is diagonal the S+L model (22) can be understood as a factor analysis model, [21], because conditional dependence relations among the manifest variables are only through the latent variables (or factors).

### C. AR Identification of Latent-variable Graphical Models

Let  $x := \begin{bmatrix} (x^m)^T & (x^l)^T \end{bmatrix}^T$  be an autoregressive process. We assume that a finite-length realization of  $x^m$  is available, i.e.  $x^m(1), x^m(2), \dots, x^m(N)$ . Regarding  $x^l$ , we have no data originated from it and its dimension  $l$  is not even known. We would compute an estimate of the spectral density  $\Phi$  of  $x$ . From the data, we can compute the  $n$ -length windowed correlogram  $\hat{\Phi}_m$  of  $x^m$ . Then, the idea is to solve the optimization problem (16) for the spectral density  $\Phi_m$  of  $x^m$  under the structural assumption (22), but not knowing in advance the supporting subspaces (23). This leads us to estimate  $\mathcal{V}_{E_m}$  and  $\mathcal{V}_G$  first, and then estimate  $\Phi_m$  consistently with the identified vector subspaces. Since the resulting estimate of the spectral density of  $x^m$  obeys to (20), it is then possible to recover the spectral density  $\Phi$  through (20) and (21).

*S+L Subspace estimation:* We propose to estimate the subspaces (23) by solving a regularized version of (16), that is,

$$\begin{aligned} (\tilde{\Sigma}, \tilde{\Lambda}) = & \arg \min_{\Sigma, \Lambda \in \mathcal{Q}_{m,n}} \int \left( -\log(\Sigma - \Lambda) + \left\langle \Sigma - \Lambda, \hat{\Phi}_m \right\rangle \right) \\ & + \lambda (\gamma \phi_1(\Sigma) + \phi_*(\Lambda)) \\ \text{subject to } & \Sigma - \Lambda \succ 0 \\ & \Lambda \succeq 0 \end{aligned} \quad (27)$$

Here,  $\lambda > 0$  and the regularizer is a combination of two penalty functions  $\phi_1$  and  $\phi_*$  inducing sparsity and low-rank on  $\Sigma$  and  $\Lambda$ , respectively. The balance between the two regularizers is



tuned by  $\gamma > 0$ . Since  $(\tilde{\Sigma} - \tilde{\Lambda})^{-1}$  represents a regularized estimate of  $\Phi_m$ ,  $\mathcal{V}_{E_m}$  is given by the support of  $\tilde{\Sigma}$  and  $\mathcal{V}_G$  by  $\tilde{\Lambda} = \Delta G^T G \Delta^*$ . Note that, for  $n = 0$ ,  $\Sigma$ ,  $\Lambda$  and  $\hat{\Phi}_m$  are matrices, i.e. the model reduces to a gaussian random vector. In this particular situation, (27) boils down to the regularization problem studied in [3] for gaussian random vectors with latent variables: in that case,  $\phi_1(\Sigma)$  is the  $\ell_1$ -norm of  $\Sigma$  and  $\phi_*(\Lambda)$  the nuclear norm of  $\Lambda$ .

*AR model identification:* For a fixed graphical model structure, that is, once the subspaces  $\mathcal{V}_{E_m}$  and  $\mathcal{V}_G$  have been identified, the optimal AR model is the solution to (16), which becomes

$$\begin{aligned}
 (\Sigma^\circ, \Lambda^\circ) = & \arg \min_{\Sigma, \Lambda \in \mathcal{Q}_{m,n}} \int \left( -\log(\Sigma - \Lambda) + \left\langle \Sigma - \Lambda, \hat{\Phi}_m \right\rangle \right) \\
 \text{subject to } & \Sigma - \Lambda \succ 0 \\
 & \Lambda \succeq 0 \\
 & \Sigma \in \mathcal{V}_{E_m} \\
 & \Lambda \in \mathcal{V}_G
 \end{aligned} \tag{28}$$

and the optimal estimate of  $\Phi_m$  is  $\hat{\Phi}_m^\circ = (\Sigma^\circ - \Lambda^\circ)^{-1}$ .

Because the identified subspaces  $\mathcal{V}_{E_m}$  and  $\mathcal{V}_G$  depend on the regularization parameters, a general identification procedure is as follows:

- i) Estimate the first  $n$  covariance lags of the manifest process as in (13)
- ii) For each  $(\lambda_k, \gamma_k)$  in a given regularization path  $\{(\lambda_k, \gamma_k)\}_{k=1}^M$ :
  - Estimate the vector subspaces  $\mathcal{V}_{E_m}$  and  $\mathcal{V}_G$
  - Compute an AR estimate  $\hat{\Phi}_m^\circ$  of  $\Phi_m$  such that  $(\hat{\Phi}_m^\circ)^{-1} \in \mathcal{V}_{E_m} + \mathcal{V}_G$
- iii) Score the identified models through a function that trades off the adherence to the data and the complexity of the models and choose the model with the minimum score
- iv) From the chosen optimal solution  $\hat{\Phi}_m^\circ = (\Sigma^\circ - \Lambda^\circ)^{-1}$ , an estimate of  $\Phi$  is

$$\hat{\Phi} = \begin{bmatrix} \hat{\Upsilon}_m & \hat{\Upsilon}_{lm}^* \\ \hat{\Upsilon}_{lm} & \hat{\Upsilon}_l \end{bmatrix}^{-1} \tag{29}$$

where  $\hat{\Upsilon}_m = \Sigma^\circ$ ,  $\hat{\Upsilon}_{lm}$  and  $\hat{\Upsilon}_l$  are such that  $\Lambda^\circ = \hat{\Upsilon}_{lm}^* \hat{\Upsilon}_l^{-1} \hat{\Upsilon}_{lm}$ .

It is worth noting that given  $\Lambda^\circ$ ,  $\hat{\Upsilon}_{lm}$  and  $\hat{\Upsilon}_l$  are known up to an  $l \times l$  invertible function. However, this is not an issue because the aim of latent variables is to explain manifest variables.

*Remark 2.1:* Since  $(\tilde{\Sigma} - \tilde{\Lambda})^{-1}$  represents a regularized estimate of  $\Phi_m$ , one would wonder why it is required to solve the second problem in order to recover an estimate of  $\Phi_m$ . As we will see in Section IV,  $\hat{\Phi}_m^\circ$  is the maximum entropy solution of a covariance extension problem. Besides such meaningful interpretation,  $\hat{\Phi}_m^\circ$  matches equality and inequality constraints imposed by the estimates  $\hat{R}_j$ 's which are reliable because typically we have  $n \ll N$ , whereas  $(\tilde{\Sigma} - \tilde{\Lambda})^{-1}$  does not.

The remainder of the paper is organized as follows: the optimization problem (27), leading to the estimation of the sparsity and low-rank subspaces  $\mathcal{V}_{E_m}$  and  $\mathcal{V}_G$ , respectively, is studied in Section III. The optimization problem (28), leading to the AR model for a fixed graphical model structure, is studied in Section IV. Finally, Section V provides an illustration of the full identification procedure.

### III. S+L SUBSPACE ESTIMATION

#### A. Primal formulation

A matrix formulation of the program (27) uses (11), which allows to parametrize  $\Sigma - \Lambda$  and  $\Lambda \in \mathcal{Q}_{m,n}$  as

$$\begin{aligned}\Sigma - \Lambda &= \Delta X \Delta^* \in \mathcal{Q}_{m,n} \\ \Lambda &= \Delta L \Delta^* \in \mathcal{Q}_{m,n}\end{aligned}\tag{30}$$

where  $X$  and  $L$  are now matrix variables in the vector space  $\mathbf{Q}_{m(n+1)}$ . Note that  $\Sigma = \Delta(X + L)\Delta^*$ . Next we reformulate (27) in terms of  $X$  and  $L$ .

1) *Positivity constraints*  $\Sigma - \Lambda \succ 0$  and  $\Lambda \succeq 0$ :

*Lemma 3.1:* Let  $\Lambda \in \mathcal{Q}_{m,n}$ . Then  $\Lambda \succeq 0$  if and only if there exists  $L \in \mathbf{Q}_{m(n+1)}$  such that  $L \succeq 0$ .

The proof is provided in Appendix A.

In view of Lemma 3.1, we replace the condition  $\Lambda \succeq 0$  with  $L \succeq 0$  and  $\Sigma - \Lambda \succ 0$  with  $X \succeq 0$ . The latter only guarantees that  $\Sigma - \Lambda \succeq 0$ . However, we will show that  $X \succeq 0$  is sufficient to guarantee that  $\Sigma - \Lambda \succ 0$  at the optimum of (27).

2) *The objective function:* Since  $\Sigma - \Lambda = \Delta X \Delta^*$  with  $X \succeq 0$ , then there exists  $A \in \mathbb{R}^{m \times m(n+1)}$  such that  $X = A^T A$ . By using *Jensen's formula*, [22, p. 184], we obtain

$$\begin{aligned} \int \log \det(\Sigma - \Lambda) &= \int \log \det(\Delta A^T A \Delta^*) \\ &= \log \det(A_0^T A_0) = \log \det X_{00}. \end{aligned}$$

Clearly, the relation above holds provided that  $X_{00} \succ 0$ . Moreover,

$$\begin{aligned} \langle \Sigma - \Lambda, \hat{\Phi}_m \rangle &= \langle \Delta X \Delta^*, \hat{\Phi}_m \rangle \\ &= \left\langle \int \Delta^* \hat{\Phi}_m \Delta, X \right\rangle = \langle T(\hat{R}), X \rangle \end{aligned}$$

where we exploited the fact that

$$\int \Delta^* \hat{\Phi}_m \Delta = T(\hat{R}). \quad (31)$$

We conclude that the objective function of (27) admits the matrix formulation

$$\begin{aligned} \int \left( -\log \det(\Sigma - \Lambda) + \langle \Sigma - \Lambda, \hat{\Phi}_m \rangle \right) \\ = -\log \det X_{00} + \langle T(\hat{R}), X \rangle. \end{aligned} \quad (32)$$

3) *The sparsity regularizer:* Let  $\Sigma \in \mathcal{Q}_{m,n}$  be such that  $\Sigma(e^{i\vartheta}) = \sum_{j=-n}^n e^{-ij\vartheta} S_j$ . Then,

$$P_{E_m^c}(\Sigma) = 0 \iff P_{E_m^c}(S_j) = 0 \quad j = 0 \dots n. \quad (33)$$

Recall that  $\Sigma = \Delta(X + L)\Delta^*$ . In view of (10), we obtain

$$P_{E_m^c}(\Sigma) = 0 \iff P_{E_m^c}(D(X + L)) = 0. \quad (34)$$

We conclude that the sparsity regularizer must induce the same sparsity on the matrices  $Y_j := D_j(X + L)$  with  $j = 0 \dots n$ . In [10], the following regularizer for  $Y \in \mathbf{M}_{m,n}$  has been proposed:

$$h_\infty(Y) = \sum_{k>h} \max \left\{ |(Y_0)_{hk}|, \max_{j=1\dots n} |(Y_j)_{hk}|, \max_{j=1\dots n} |(Y_j)_{kh}| \right\}. \quad (35)$$

Let  $v_{kh}$ , with  $k > h$ , be the vector of  $(k, h)$  and  $(h, k)$  entries of the coefficients  $Y_j$  with  $j = 0 \dots n$ . Therefore,

$$h_\infty(Y) = \sum_{k>h} \|v_{kh}\|_\infty \quad (36)$$

where  $\|\cdot\|_\infty$  denotes the  $\ell_\infty$ -norm. On the other hand,  $h_\infty(Y)$  is the  $\ell_1$ -norm of the vector having (nonnegative) entries  $\|v_{kh}\|_\infty$  with  $k > h$ . Accordingly,  $h_\infty(Y)$  encourages sparsity among  $v_{kh}$ 's, that is induces the same sparsity on the matrices  $Y_j$   $j = 0 \dots n$ .

4) *The low-rank regularizer:*

*Proposition 3.1:* Given  $\Lambda \in \mathcal{A}_m$ , we define the convex function

$$\phi_*(\Lambda) := \sum_{k=1}^m \int \sigma_k(\Lambda) \quad (37)$$

and the restricted rank function

$$\text{rank}'(\Lambda) := \begin{cases} \text{rank}(\Lambda), & \|\Lambda\| \leq 1 \\ +\infty, & \text{otherwise.} \end{cases} \quad (38)$$

Then, the convex hull of  $\text{rank}'(\Lambda)$  is

$$\begin{cases} \phi_*(\Lambda), & \|\Lambda\| \leq 1 \\ +\infty, & \text{otherwise.} \end{cases} \quad (39)$$

The proof is provided in Appendix B.

We conclude that  $\phi_*(\Lambda)$  defined in (37) is the adequate regularizer of  $\text{rank}(\Lambda)$ . Since  $\Lambda \succeq 0$ ,  $\sigma_k(\Lambda(e^{i\vartheta}))$  represents the  $k$ -th eigenvalue of  $\Lambda(e^{i\vartheta})$ . Thus,  $\phi_*(\Lambda) = \text{tr} \int \Lambda$ . Finally,

$$\phi_*(\Lambda) = \text{tr} \int \Delta L \Delta^* = \text{tr} \left( L \int \Delta^* \Delta \right) = \text{tr}(L)$$

where we exploited the fact that

$$\int e^{ij\vartheta} = \begin{cases} 1, & j = 0 \\ 0, & j \neq 0. \end{cases} \quad (40)$$

5) *Primal Formulation:* By collecting the results in 1)-4), we rewrite (27) as

$$\begin{aligned} (X^\circ, L^\circ) = \arg \min_{X, L \in \mathbf{Q}_{m(n+1)}} & -\log \det X_{00} + \left\langle \mathbf{T}(\hat{R}), X \right\rangle \\ & + \lambda \gamma h_\infty(\mathbf{D}(X + L)) + \lambda \text{tr}(L) \\ \text{subject to} & X_{00} \succ 0, X \succeq 0, L \succeq 0 \end{aligned} \quad (41)$$

Formulation (41) and (27) are equivalent provided that  $\Delta X^\circ \Delta^* \succ 0$ . Finally, it is worth noting that (41) is a generalization of the regularized problem studied in [10]. The problem formulations coincide when  $L = 0$ , that is for estimating an AR process having a sparse graphical model but no latent variables.

### B. Dual formulation

We show that (41) does admit a solution by exploiting duality theory. First, note that (41) is strictly feasible (pick  $X = I$  and  $L = I$ ), thus *Slater's condition* holds. Accordingly, the duality gap between (41) and its dual problem is equal to zero. We introduce a new variable  $Y \in \mathbf{M}_{m,n}$  in (41) to obtain the following equivalent problem

$$\begin{aligned} & \arg \min_{\substack{X, L \in \mathbf{Q}_{m(n+1)} \\ Y \in \mathbf{M}_{m,n}}} -\log \det X_{00} + \left\langle T(\hat{R}), X \right\rangle + \lambda \gamma h_{\infty}(Y) + \lambda \operatorname{tr}(L) \\ & \text{subject to} \quad X_{00} \succ 0, \quad X \succeq 0, \quad L \succeq 0 \\ & \quad \quad \quad Y = D(X + L) \end{aligned}$$

The Lagrangian is

$$\begin{aligned} \mathcal{L}(X, L, Y, U, V, Z) &= -\log \det X_{00} + \left\langle T(\hat{R}), X \right\rangle + \lambda \gamma h_{\infty}(Y) + \lambda \operatorname{tr}(L) \\ &\quad - \langle U, X \rangle - \langle V, L \rangle + \langle Z, D(X + L) - Y \rangle \\ &= -\log \det X_{00} + \left\langle T(\hat{R}) - U, X \right\rangle + \langle \lambda I - V, L \rangle \\ &\quad + \lambda \gamma h_{\infty}(Y) - \langle Z, Y \rangle + \langle T(Z), X + L \rangle \\ &= -\log \det X_{00} + \left\langle T(\hat{R}) + T(Z) - U, X \right\rangle \\ &\quad + \langle \lambda I + T(Z) - V, L \rangle + \lambda \gamma h_{\infty}(Y) - \langle Z, Y \rangle \end{aligned}$$

where  $U, V \in \mathbf{Q}_{m(n+1)}$  such that  $U, V \succeq 0$  and  $Z \in \mathbf{M}_{m,n}$ . The dual function is the infimum of  $\mathcal{L}$  over  $X, L$  and  $Y$ . We start by minimizing with respect to  $Y$ . The Lagrangian  $\mathcal{L}$  depends on  $Y$  only through the term

$$\lambda \gamma h_{\infty}(Y) - \langle Z, Y \rangle, \quad (42)$$

which was shown, [10], to be bounded below only if

$$\operatorname{diag}(Z_j) = 0, \quad j = 0 \dots n \quad (43)$$

$$\sum_{j=0}^n |(Z_j)_{kh}| + |(Z_j)_{hk}| \leq \lambda \gamma, \quad k \neq h, \quad (44)$$

in which case the infimum is equal to zero. The partial minimization of the Lagrangian over  $Y$  is therefore

$$\inf_Y \mathcal{L} = \begin{cases} -\log \det X_{00} + \langle T(\hat{R}) + T(Z) - U, X \rangle \\ \quad + \langle \lambda I + T(Z) - V, L \rangle & (43), (44) \\ -\infty & \text{otherwise.} \end{cases}$$

Likewise, the Lagrangian  $\mathcal{L}$  depends on  $L$  only through the term  $\langle \lambda I + T(Z) - V, L \rangle$ , which is bounded below only if

$$\lambda I + T(Z) - V = 0, \quad (45)$$

in which case the infimum is equal to zero. Thus,

$$\inf_{L,Y} \mathcal{L} = \begin{cases} -\log \det X_{00} + \langle T(\hat{R}) + T(Z) - U, X \rangle & (43)-(45) \\ -\infty & \text{otherwise.} \end{cases}$$

Finally, the terms in  $X_{00}$  are bounded below if and only if

$$(T(Z) + T(\hat{R}) - U)_{00} \succ 0 \quad (46)$$

and if (46) holds, they are minimized by  $X_{00} = (T(Z) + T(\hat{R}) - U)_{00}^{-1}$ . The Lagrangian is linear in the remaining variables  $X_{kh}$ , and therefore bounded below (and identically zero) only if

$$(T(Z) + T(\hat{R}) - U)_{kh} = 0 \quad \forall (k, h) \neq (0, 0). \quad (47)$$

The final expression for the dual functional is

$$\inf_{X,L,Y} \mathcal{L} = \begin{cases} \log \det(T(Z) + T(\hat{R}) - U)_{00} + m & (43)-(47) \\ -\infty & \text{otherwise.} \end{cases} \quad (48)$$

The dual problem consists in maximizing the dual functional (48) with respect to  $U$ ,  $V$  and  $Z$  subject to the constraints  $U \succeq 0$  and  $V \succeq 0$ . Moreover, eliminating the slack variables  $U$  and

$V$ , and adding the variable  $W := (T(Z) + T(\hat{R}) - U)_{00}$  the dual problem takes the final form

$$\begin{aligned}
& \max_{\substack{W \in \mathbf{Q}_m \\ Z \in \mathbf{M}_{m,n}}} \log \det W + m \\
& \text{subject to } W \succ 0 \\
& T(\hat{R}) + T(Z) \succeq \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix} \\
& \text{diag}(Z_j) = 0, \quad j = 0 \dots n \\
& \sum_{j=0}^n |(Z_j)_{kh}| + |(Z_j)_{hk}| \leq \lambda \gamma, \quad k \neq h \\
& \lambda I + T(Z) \succeq 0
\end{aligned} \tag{49}$$

*Proposition 3.2:* Problem (49) admits a solution.

The proof is provided in Appendix C.

From the next statement we conclude that Problem (27) admits a solution.

*Proposition 3.3:* Problem (41) admits a solution  $(X^\circ, L^\circ)$  such that  $\Delta X^\circ \Delta^* \succ 0$ . Accordingly (27) and (41) are equivalent. Moreover,  $X^\circ$  is unique.

The proof is provided in Appendix D.

It is worth noting that (49) is easier to solve than (41), because the objective function in (49) is smooth.

### C. Estimation of the Vector Subspaces

The vector subspace  $\mathcal{V}_{E_m}$  is given by the support of  $\tilde{\Sigma} = \Delta(X^\circ + L^\circ)\Delta^*$ . In view of (10), we obtain

$$E_m^c = \{(k, h) \in V_m \times V_m \text{ s.t. } (D(X^\circ + L^\circ))_{kh} = 0\} \tag{50}$$

and hence also  $\mathcal{V}_{E_m}$ . Since  $\tilde{\Lambda} = \Delta L^\circ \Delta^*$ , the vector subspace  $\mathcal{V}_G$  is the column space of  $L^\circ$ , given by the decomposition  $L^\circ = G^T G$  where  $G$  is a full row rank matrix.

Next, we show how to recover  $(X^\circ, L^\circ)$  from an optimal solution  $(W^\circ, Z^\circ)$  of the smooth convex optimization program (49). Such a recovering scheme also provides sufficient conditions for the uniqueness of the two vector subspaces. Regarding  $X^\circ$ , let  $B \in \mathbb{R}^{m \times m(n+1)}$  the solution

of the *Yule-Walker* equation

$$(\mathbf{T}(\hat{R}) + \mathbf{T}(Z^\circ))B^T = \begin{bmatrix} W^\circ \\ 0 \end{bmatrix}, \quad B_0 = I \quad (51)$$

then  $X^\circ = B^T(W^\circ)^{-1}B$ , see Appendix D for more details. Next, we deal with the recovering of  $L^\circ$ . Because of the strong duality between (41) and (49), we have

$$\langle V^\circ, L^\circ \rangle = 0 \quad (52)$$

where  $V^\circ := \lambda I + \mathbf{T}(Z^\circ)$ , see (45). If  $V^\circ$  is a full rank matrix then, in view of (52),  $L^\circ = 0$  is the unique solution,  $\mathcal{V}_G = \{0\}$  and  $\mathcal{V}_{E_m}$  is univocally characterized by (50). Otherwise, let  $l > 0$  be the dimension of the nullspace of  $V^\circ$ . Then there exists a full row rank matrix  $G \in \mathbb{R}^{l \times m(n+1)}$  such that  $V^\circ G^T = 0$ . Since  $V^\circ, L^\circ \succeq 0$ , from (52) it follows that

$$L^\circ = G^T H G \quad (53)$$

where  $H$ , unknown, belongs to  $\mathbf{Q}_l$  and  $H \succeq 0$ . Therefore,  $L^\circ$  is known up to the (scaling) factor  $H$ . The minimization of (42) under constraints (43) and (44) is equivalent to the minimization of the non-negative function

$$\begin{aligned} & \max\{|(Y_0)_{kh}|, \max_{j=1 \dots n} |(Y_j)_{kh}|, \max_{j=1 \dots n} |(Y_j)_{hk}|\} \\ & \times \left( \lambda\gamma - \sum_{j=0}^n |(Z_j)_{kh}| + |(Z_j)_{hk}| \right), \end{aligned} \quad (54)$$

for each  $k > h$ , subject to the constraint that their sum is bounded by  $\lambda\gamma h_\infty(Y) - \langle Z, Y \rangle$ . Since the optimal value of (42) is always equal to zero, then the optimal value of (54) is equal to zero for each  $k > h$ . Thus, if  $\sum_{j=0}^n |(Z_j)_{kh}| + |(Z_j)_{hk}| < \lambda\gamma$  then

$$\max\{|(Y_0)_{kh}|, \max_{j=1 \dots n} |(Y_j)_{kh}|, \max_{j=1 \dots n} |(Y_j)_{hk}|\} = 0 \quad (55)$$

and  $(Y_j)_{kh} = (Y_j)_{hk} = 0$  with  $j = 0 \dots n$ . Since  $Y = \mathbf{D}(X + L)$ ,  $\sum_{j=0}^n |(Z_j)_{kh}| + |(Z_j)_{hk}| < \lambda\gamma$  implies that  $(\mathbf{D}_j(X + L))_{kh} = (\mathbf{D}_j(X + L))_{hk} = 0$  with  $j = 0 \dots n$ . Accordingly,  $H$  is obtained by solving the following system of linear equations

$$(\mathbf{D}_j(X^\circ + G^T H G))_{kh} = 0 \quad j = 0 \dots n, \quad \forall (k, h) \in I. \quad (56)$$

where

$$I := \left\{ (k, h) \text{ s.t. } k \neq h, \sum_{j=0}^n |(Z_j)_{kh}| + |(Z_j)_{hk}| < \lambda\gamma \right\}. \quad (57)$$



Note that (56) is a system of  $(n+1) \times |I|$  equations with  $l(l+1)/2$  unknowns (i.e. the number of independent parameters in  $H$ ). For  $\lambda\gamma$  and  $\gamma$  sufficiently large,  $|I|$  would be sufficiently large and  $l$  sufficiently small, respectively, so that (56) admits a unique solution. We stress the fact that it may happen that (56) has not unique solution even  $l(l+1)/2 \ll (n+1) \times |I|$ . As observed in [3], this is more likely when  $\mathcal{V}_G$  contains sparse elements, that is the latent variables are not sufficiently “diffuse” across the manifest variables, or  $\mathcal{V}_{E_m}$  contains elements with a low degree of sparsity, that is the are manifest variables conditionally dependent to too many other manifest variables. Both cases may lead to a non-identifiability of the AR model solution to Problem (28) because some sparse and low-rank components are not distinguishable. One avoids those situations checking that (56) has unique solution. We formalize the above observation.

*Proposition 3.4:* If (56) admits a unique solution, then  $\mathcal{V}_{E_m}$  and  $\mathcal{V}_G$  are unique and have *transverse intersection*, i.e.  $\mathcal{V}_{E_m} \cap \mathcal{V}_G = \{0\}$ .

The proof is provided in Appendix E.

The transversality condition means that any element of  $\mathcal{V}_{E_m} + \mathcal{V}_G$  admits a unique decomposition into the two subspaces. We will see in Section IV that this condition guarantees the uniqueness of the solution to Problem (28).

#### IV. AR MODEL IDENTIFICATION

The convex formulation of the convex optimization Problem (28) parallels the developments in the previous section. We adopt the parametrization

$$\begin{aligned}\Sigma - \Lambda &= \Delta X \Delta^* \\ \Lambda &= \Delta L \Delta^* = \Delta G^T H G \Delta^*\end{aligned}\tag{58}$$

where the matrix unknowns are  $X \in \mathbf{Q}_{m(n+1)}$  and  $H \in \mathbf{Q}_l$ . Note that  $\Sigma = \Delta(X + G^T H G)\Delta^*$ . The positivity conditions  $\Sigma - \Lambda \succ 0$  and  $\Lambda \succeq 0$  are replaced by  $X \succeq 0$  and  $H \succeq 0$ , respectively. Also in this case  $X \succeq 0$  only guarantees that  $\Sigma - \Lambda \succeq 0$ . In view of (34), condition  $\Sigma \in \mathcal{V}_{E_m}$  is replaced by  $P_{E_m^c}(D(X + G^T H G)) = 0$ . Clearly condition  $\Lambda \in \mathcal{V}_G$  follows from the chosen parametrization. Finally, the objective function is given by (32) provided that  $X_{00} \succ 0$ . The

convex program (28) thus admits the matrix formulation

$$\begin{aligned}
& \min_{\substack{X \in \mathbf{Q}_{m(n+1)} \\ H \in \mathbf{Q}_l}} -\log \det X_{00} + \left\langle \mathbf{T}(\hat{R}), X \right\rangle \\
& \text{subject to } X_{00} \succ 0, X \succeq 0, H \succeq 0 \\
& P_{E_m^c}(\mathbf{D}(X + G^T H G)) = 0
\end{aligned} \tag{59}$$

Both formulations are equivalent provided that the optimal solution, say  $(X^\circ, H^\circ)$ , is such that  $\Delta X^\circ \Delta^* \succ 0$ .

*Proposition 4.1:* Problem (59) does admit a solution. Moreover,  $\Delta X^\circ \Delta$  is unique and such that  $\Delta X^\circ \Delta \succ 0$ .

The proof is provided in Appendix F.

The optimal spectral density  $\hat{\Phi}_m^\circ$  thus admits the matrix decomposition

$$(\hat{\Phi}_m^\circ)^{-1} = \Delta X^\circ \Delta^* = \underbrace{\Delta(X^\circ + G^T H^\circ G) \Delta^*}_{\in \mathcal{V}_{E_m}} - \underbrace{\Delta G^T H^\circ G \Delta^*}_{\in \mathcal{V}_G} \tag{60}$$

which is unique when  $\mathcal{V}_{E_m}$  and  $\mathcal{V}_G$  have transverse intersection.

*Corollary 4.1:* The AR latent-variable graphical model solution to Problem (28) is unique when  $\mathcal{V}_{E_m}$  and  $\mathcal{V}_G$  are estimated from (27) with  $\lambda\gamma$  and  $\lambda$  sufficiently large.

We now give an important interpretation of the optimal solution of (28). Consider the following covariance extension problem.

*Problem 1:* Find  $\Phi_m \in \mathcal{S}_m$  such that

$$\begin{aligned}
& P_{E_m} \left( \int \Delta \Phi_m - \hat{R} \right) = 0 \\
& \int G \Delta^* \Phi_m \Delta G^T \succeq G \mathbf{T}(\hat{R}) G^T.
\end{aligned} \tag{61}$$

The condition

$$\int \Delta \Phi_m = \hat{R} \tag{62}$$

implies that  $P_{E_m} \left( \int \Delta \Phi_m - \hat{R} \right) = 0$ . Moreover, (62) is equivalent to  $\int \Delta^* \Phi_m \Delta = \mathbf{T}(\hat{R})$  which implies that  $\int G \Delta^* \Phi_m \Delta G^T \succeq G \mathbf{T}(\hat{R}) G^T$ . Accordingly, Problem 1 is a relaxation of the classic covariance extension problem. The next theorem shows that  $\hat{\Phi}_m^\circ$  is the maximum entropy solution of Problem 1.

*Theorem 4.1:* Problem (28) is the dual of the convex optimization problem

$$\begin{aligned} & \max_{\Phi_m \in \mathcal{S}_m} \int \log \det \Phi_m \\ & \text{subject to } P_{E_m} \left( \int \Delta \Phi_m - \hat{R} \right) = 0 \\ & \int G \Delta^* \Phi_m \Delta G^T \succeq G T(\hat{R}) G^T \end{aligned} \quad (63)$$

*Proof:* Note that, (63) is a relaxation of (14). Moreover, (14) admits solution (and thus it is feasible), because  $T(\hat{R}) \succ 0$ . Accordingly, (63) is feasible. Moreover, we only have linear inequality constraints in (63) which implies the *refined Slater's condition* [23]. Thus we have strong duality for (63) and its dual. The Lagrange functional is:

$$\begin{aligned} \mathcal{L}(\Phi_m, S, H) &= \int \log \det \Phi_m - \left\langle P_{E_m} \left( \int \Delta \Phi_m - \hat{R} \right), S \right\rangle \\ &+ \left\langle \int G \Delta^* \Phi_m \Delta G^T - G T(\hat{R}) G^T, H \right\rangle \end{aligned} \quad (64)$$

where  $H \in \mathbf{Q}_l$  such that  $H \succeq 0$ , and  $S \in \mathbf{M}_{m,n}$ . Moreover,

$$\begin{aligned} \mathcal{L}(\Phi_m, S, H) &= \int \log \det \Phi_m - \left\langle \int \Delta \Phi_m - \hat{R}, P_{E_m}(S) \right\rangle \\ &+ \left\langle \int G \Delta^* (\Phi_m - \hat{\Phi}_m) \Delta G^T, H \right\rangle \\ &= \int \log \det \Phi_m - \left\langle \int \Delta \Phi_m - \hat{R}, P_{E_m}(S) \right\rangle \\ &+ \left\langle \Phi_m - \hat{\Phi}_m, \Delta G^T H G \Delta^* \right\rangle \end{aligned} \quad (65)$$

where we exploited (31) and the fact that  $P_{E_m}$  is a self-adjoint operator. By defining  $\Sigma := P_{E_m}(S_0) + \frac{1}{2} \sum_{j=1}^n e^{-ij\vartheta} P_{E_m}(S_j) + e^{ij\vartheta} P_{E_m}(S_j)^T \in \mathcal{V}_{E_m}$  and  $\Lambda := \Delta G^T H G \Delta^* \in \mathcal{V}_G$  such that  $\Lambda \succeq 0$ , we obtain the following compact notation for the Lagrangian

$$\begin{aligned} \mathcal{L}(\Phi_m, \Sigma, \Lambda) &= \int \left( \log \det \Phi_m - \left\langle \Phi_m - \hat{\Phi}_m, \Sigma \right\rangle + \left\langle \Phi_m - \hat{\Phi}_m, \Lambda \right\rangle \right). \end{aligned}$$

Since  $\mathcal{L}(\cdot, \Sigma, \Lambda)$  is strictly concave over  $\mathcal{S}_m$ , its unique maximum point is given by annihilating its first variation in each direction  $\delta \Phi_m \in \mathbf{L}_2^{m \times m}$ :

$$\delta \mathcal{L}(\Phi_m, \Sigma, \Lambda; \delta \Phi_m) = \text{tr} \int ((\Phi_m^{-1} - \Sigma + \Lambda) \delta \Phi_m) \quad (66)$$

Note that  $\Phi_m^{-1} - \Sigma + \Lambda \in \mathbf{L}_2^{m \times m}$ , thus the first variation is zero for each  $\delta\Phi_m$  if and only if  $\Phi_m^{-1} - \Sigma + \Lambda = 0$ . Accordingly, if  $\Sigma - \Lambda \succ 0$  then the unique maximum point of  $\mathcal{L}(\cdot, \Sigma, \Lambda)$  is

$$\hat{\Phi}_m^\circ := (\Sigma - \Lambda)^{-1} \quad (67)$$

with  $\Sigma \in \mathcal{V}_{E_m}$  and  $\Lambda \in \mathcal{V}_G$  such that  $\Lambda \succeq 0$ . Then, by substituting (67) in the Lagrangian we obtain, up to a constant term, the objective function of (28). ■

The interpretation of the convex program (28) as the dual of a covariance extension problem is insightful. First, it coincides with the problem considered in [12] for the AR case, since the solution satisfies  $G = 0$  when the inequality constraint in (63) is removed. On the other hand, it is worth noting that [12] considers ARMA models which are more general than the AR ones. Second, both constraints in (63) have a clear interpretation: the equality constraint imposes that the optimal spectral density  $\hat{\Phi}_m^\circ$  matches the estimated covariance lags  $\hat{R}_0 \dots \hat{R}_n$  in the positions specified by  $E_m$ . Regarding the inequality constraint, consider the stochastic process

$$y(t) = \sum_{j=0}^n G_j x^m(t-j) \quad (68)$$

whose variables are linear combinations of the  $m$  manifest variables in a time window of length  $n$ . Accordingly,  $y$  encodes information about  $x^m$ . It is readily checked that

$$\mathbb{E}[y(t)y(t)^T] = \sum_{k,h=0}^n G_k R_{h-k} G_h^T = G\mathbf{T}(R)G^T. \quad (69)$$

The inequality constraint therefore imposes that the covariance matrix of  $y$  is lower bounded by the one estimated from the data, i.e.  $G\mathbf{T}(\hat{R})G^T$ .

## V. NUMERICAL EXAMPLES

### A. Synthetic Example

We consider an AR latent-variable graphical model of order  $n = 1$  with  $m = 15$  manifest variables,  $l = 1$  latent variable. Its interaction graph is depicted in Figure 2(a). We generate a data sequence of length  $N = 500$  for the manifest process and we apply the identification procedure outlined at the end of Section II-C. In Figure 2(b) we depict the latent-variable graphical models obtained for different values of  $\lambda$  and  $\lambda\gamma$ . Not surprisingly, increasing the rank regularization parameter  $\lambda$  favors few latent variables, whereas by increasing the sparsity

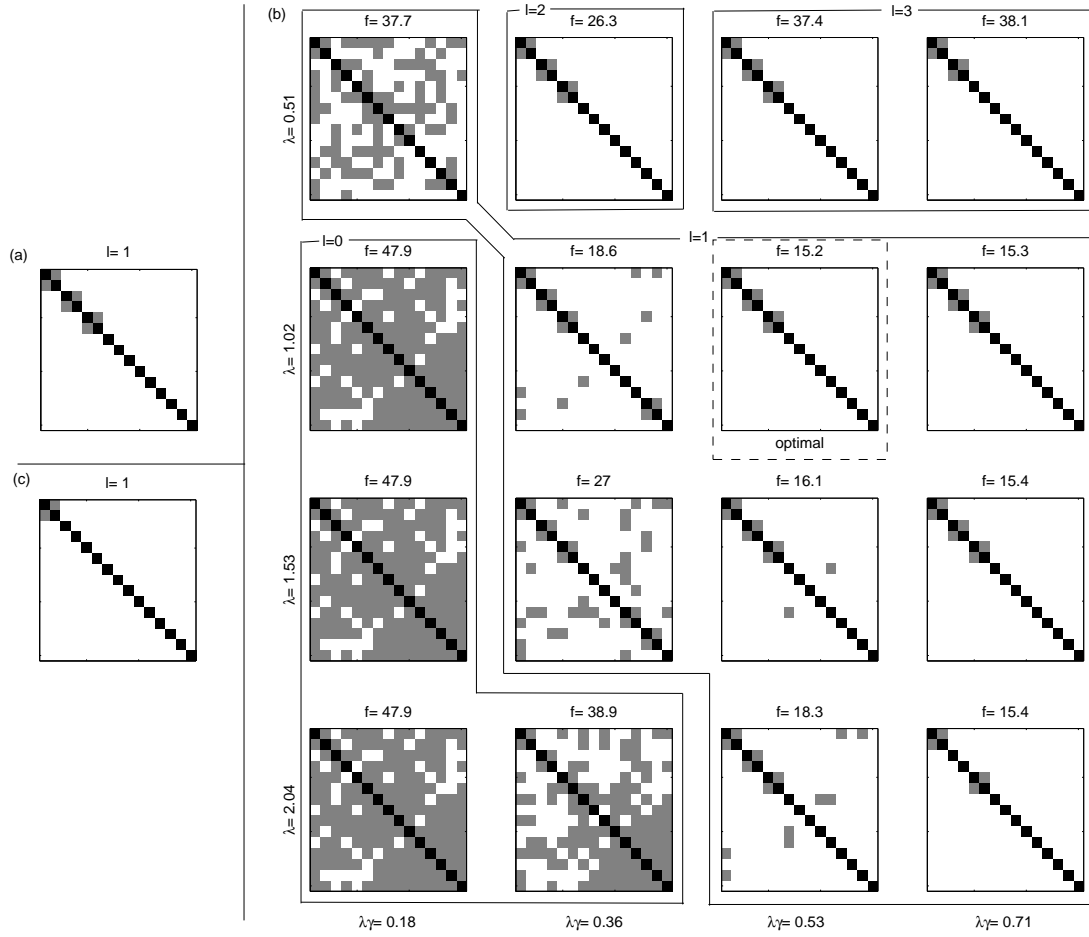


Fig. 2. (a) Interaction graph of the generated model. (b) Interaction graphs of optimal models estimated for  $n = 1$  and for different values of  $\lambda$  and  $\lambda\gamma$ . (c) Interaction graph of the optimal model estimated with  $n = 0$ . Each figure shows the interaction graph for the manifest variables: grey denotes an edge, white denotes no edge, and black denotes a manifest node. The number of latent variables and the value of the score function is indicated on the top of each figure.

regularization parameter  $\lambda\gamma$  favors few conditional dependence relations among the manifest variables.

To discriminate among models, we consider the following score function:

$$f(E_m, l, \hat{\Phi}_m^\circ, \hat{\Phi}_C) = \mathbb{D}(\hat{\Phi}_C \| \hat{\Phi}_m^\circ) \times p. \quad (70)$$

Here,  $\hat{\Phi}_C$  is the smoothed *correlogram* of  $x^m$  computed from the data by using the *Bartlett*

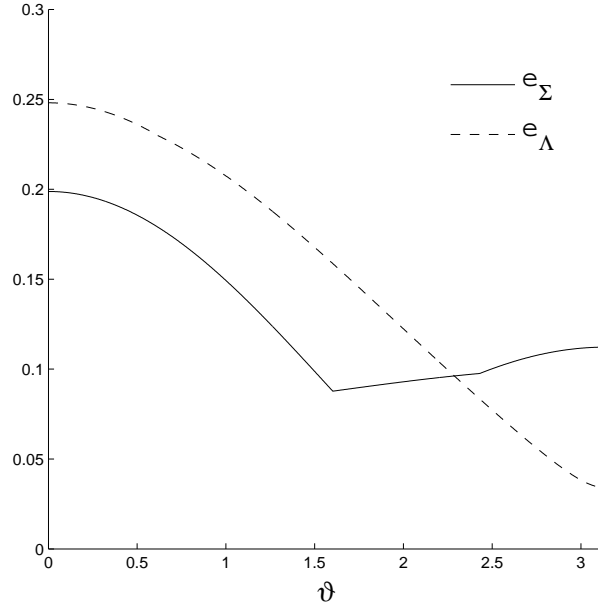


Fig. 3. Normalized estimation errors  $e_{\Sigma}(e^{i\vartheta}) = \frac{\|\Sigma(e^{i\vartheta}) - \Sigma^{\circ}(e^{i\vartheta})\|_2}{\|\Sigma\|}$  and  $e_{\Lambda}(e^{i\vartheta}) = \frac{\|\Lambda(e^{i\vartheta}) - \Lambda^{\circ}(e^{i\vartheta})\|_2}{\|\Lambda\|}$  as a function of  $\vartheta \in [0, \pi]$  for the data set in Section V.

window, [14]. The cost

$$\begin{aligned} \mathbb{D}(\hat{\Phi}_C \| \hat{\Phi}_m^{\circ}) &:= \frac{1}{2} \left( \int \left( \log \det(\hat{\Phi}_C^{-1} \hat{\Phi}_m^{\circ}) \right) \right. \\ &\quad \left. + \left\langle \hat{\Phi}_C, (\hat{\Phi}_m^{\circ})^{-1} \right\rangle - m \right) \end{aligned} \quad (71)$$

is the *relative entropy rate*, [13], between  $\hat{\Phi}_C$  and  $\hat{\Phi}_m^{\circ}$ . Thus, it ranks the adherence of  $\hat{\Phi}_m^{\circ}$  to the data. The term

$$p = (|E_m| - m) + ml \quad (72)$$

is the total number of edges in the latent-variable graphical model. Thus,  $p$  places a penalty on models with high complexity. An alternative choice for the score function would be  $\mathbb{D}(\hat{\Phi}_C \| \hat{\Phi}_m^{\circ}) + \alpha(N)p$  where the weighting  $\alpha(N)$  is the trade-off parameter between the adherence to the data and the complexity of the model. Typically  $\alpha(N)$  is a decreasing function in  $N$  because the data should reveal the simple structure as  $N$  increases. The authors of [11] recommend the choices  $\alpha(N) = N^{-1}$  and  $\alpha(N) = N/\log N$ . In contrast, the authors of [12] recommend the score function (70) because it is robust to scaling. Based on (70), the minimum value of  $f$  is equal

to 15,2 reached with  $\lambda = 1.02$  and  $\lambda\gamma = 0.53$ . Its interaction graph coincides with the true one. Figure 3 provides a graph of the normalized estimation errors of  $\Sigma$  and  $\Lambda$  at each frequency:

$$\begin{aligned} e_{\Sigma}(e^{i\vartheta}) &= \frac{\|\Sigma(e^{i\vartheta}) - \Sigma^{\circ}(e^{i\vartheta})\|_2}{\|\Sigma\|} \\ e_{\Lambda}(e^{i\vartheta}) &= \frac{\|\Lambda(e^{i\vartheta}) - \Lambda^{\circ}(e^{i\vartheta})\|_2}{\|\Lambda\|}. \end{aligned} \quad (73)$$

We found similar results by varying the sample data. Finally, we applied the same identification procedure with  $n = 0$ , i.e. by estimating a gaussian random vector. The estimated interaction graph in Figure 2(c) does not recover the generated model. This suggests the potential benefit of AR modeling in the estimation of latent-variable graphical models.

### B. International Stock Markets

The data used in this simulation consists of a time series of daily stock markets indices at closing time, in terms of local currency units, of twenty-two financial markets. The twenty-two countries and their respective price indices are: Australia (All Ordinaries index denoted AU), New Zealand (50 Gross index denoted NZ), Singapore (STI index denoted SG), Hong Kong (Hang Seng index denoted HK), China (SSE Composite index denoted CH), Japan (Nikkei225 index denoted JA), Korea (KOSPI Composite index denoted KO), Taiwan (Weighted index denoted TA), Brazil (IBOVESPA index denoted BR), Mexico (IPC index denoted ME), Argentina (Merval index denoted AR), Swiss (SMI index denoted SW), Greece (Athen Composite index denoted GR), Belgium (BFX index denoted BE), Austria (ATX index denoted AS), Germany (DAX index denoted GE), France (CAC 40 index denoted FR), Netherlands (AEX index denoted NL), United Kingdom (FTSE 100 index denoted UK), United States (S&P500 denoted US), Canada (S&PTSX Composite index denoted CA) and Malaysia (KLCI index denoted MA). The data are obtained from the website at <http://finance.yahoo.com/>. The sample period is from 4th January 2012 up to 31th December 2013. For each index, we compute the return between the trading day  $t - 1$  and  $t$  as log differences  $r_t = 100(\log p_t - \log p_{t-1})$  with  $p_t$  closing price on day  $t$ . In cases of national holidays in some country, the missing index value is replaced by the last trading day's value, that is the return is zero. The obtained data sequence has length  $N = 518$ .

We applied the identification procedure of Section V-A with  $n = 1$ . In Figure 4(b) we depict the estimated graphical model from the financial stock returns data. We found one latent variable

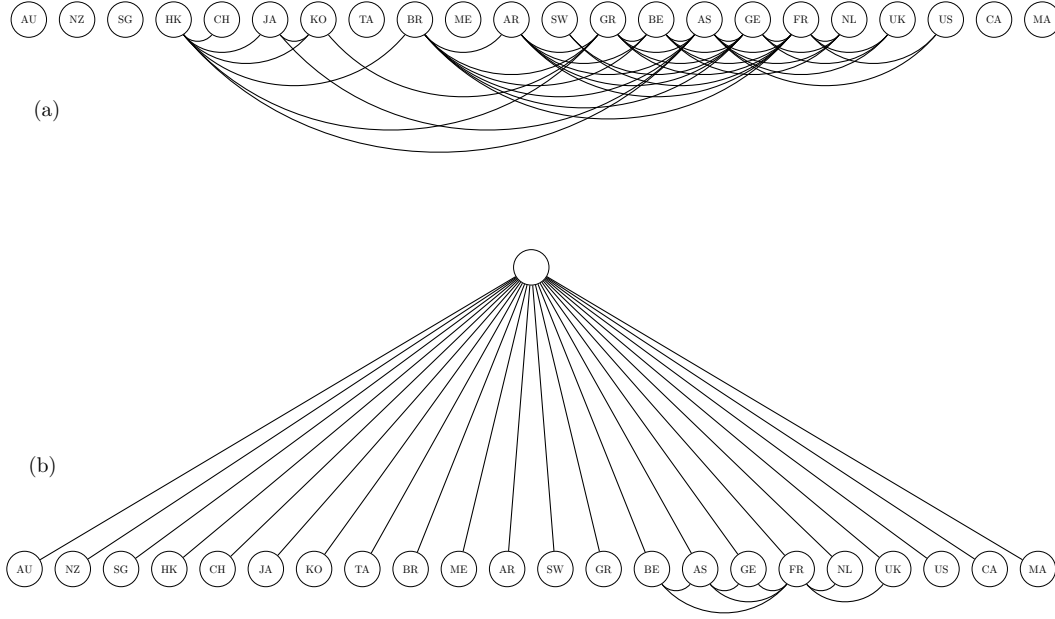


Fig. 4. Graphical models for the international financial stock returns data: (a) Best model without latent variables (b) Best model allowing latent variables.

and the total number of edges is equal to 29. It is interesting to observe that the latent variable is not sufficient to characterize the conditional dependence relations of Europeans markets (with exception of Greece) and the identification procedure added edges connecting them. This can be explained by the commencement of the economic and monetary union, see [24]. In Figure 4(a) we depict the estimated graphical model without latent variables which is characterized by 49 edges among the markets. It is clear that its interpretation is less intuitive than the one with the latent variable. Finally, it is worth noting that  $\mathbb{D}(\hat{\Phi}_C || \hat{\Phi}_m^\circ) \cong 3.9$  for both models, that is both models have the same adherence degree to the data.

We consider the estimated joint spectral density  $\hat{\Phi}$  of the manifest and latent variables in (29) where we choose  $\hat{\Upsilon}_l = 1$ . Its partial coherence is defined as

$$\tilde{\Phi}^{-1} = \text{diag}(\hat{\Phi}^{-1})^{-1/2} \hat{\Phi}^{-1} \text{diag}(\hat{\Phi}^{-1})^{-1/2}. \quad (74)$$

Its entry in position  $(k, h)$  represents a measure of how dependent  $x_k$  and  $x_h$  are conditioned to



$\mathcal{X}_{V_{m+l} \setminus \{k,h\}}$ . We partition the partial coherence as follows

$$\tilde{\Phi}^{-1} = \begin{bmatrix} \tilde{\Upsilon}_m & \tilde{\Upsilon}_{lm}^* \\ \tilde{\Upsilon}_{lm} & 1 \end{bmatrix}. \quad (75)$$

In Figure 4 the entries of  $\tilde{\Upsilon}_{lm}$ , representing a measure of the conditional dependence between the latent variable and the stock returns, are depicted.

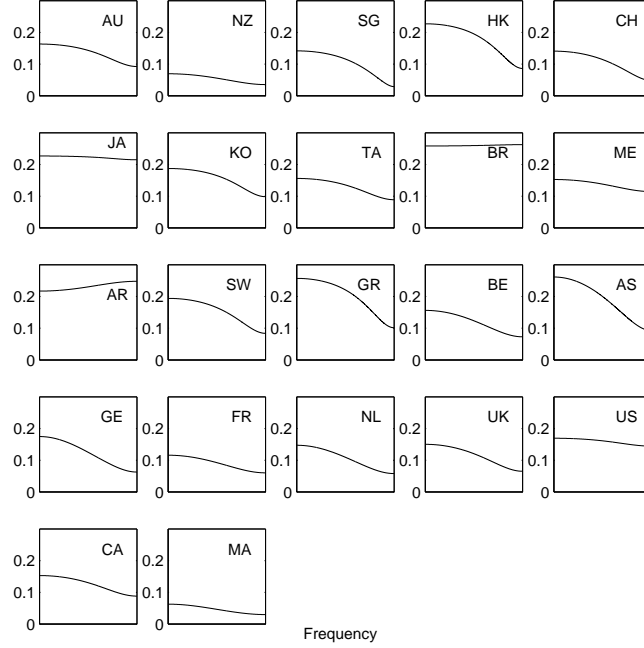


Fig. 5. Partial coherence between the latent variable and the stock returns.

## VI. CONCLUSIONS

In this paper we dealt with the identification of AR latent-variable graphical models. The inverse of the manifest spectral density of these models admits a sparse plus low-rank decomposition, captured in two distinct vector subspaces. We presented a two-step procedure for estimating such models. A first optimization problem uses sparsity and low-rank regularizers to estimate the two vector subspaces. A second optimization problem performs the AR identification restricted to those vector subspaces. Through duality, the second problem provides a novel covariance extension problem. We provided a simulation study to illustrate the proposed methodology.

Finally, we tested our method to international stock return data where the introduction of a latent variable led to a simple graphical model without compromising the adherence to the data.

## APPENDIX

### A. Proof of Lemma 3.1

If  $L \succeq 0$ , then there exists  $C$  such that  $L = CC^T$ . Accordingly,  $\Delta(e^{i\vartheta})L\Delta(e^{i\vartheta})^* = (\Delta(e^{i\vartheta})C)(\Delta(e^{i\vartheta})C)^*$  which is positive semi-definite for each  $\vartheta \in [-\pi, \pi]$ . Thus,  $\Lambda = \Delta L \Delta^* \succeq 0$ . Conversely, if  $\Lambda \in \mathcal{Q}_{m,n}$  is such that  $\Lambda \succeq 0$ , then it admits the spectral factorization  $\Lambda = \Gamma \Gamma^*$  where  $\Gamma = \Delta A^T$  such that  $A \in \mathbb{R}^{m \times m(n+1)}$ , [14]. Hence,  $\Lambda = \Delta A^T A \Delta^*$ . We conclude that  $\Lambda = \Delta L \Delta^*$  with  $L = A^T A \succeq 0$ .  $\blacksquare$

### B. Proof of Proposition 3.1

Consider an extended-real valued functional  $f : \mathcal{A}_m \rightarrow [-\infty, +\infty]$ . Its *conjugate*  $f^* : \mathcal{A}_m \rightarrow [-\infty, +\infty]$  is defined as

$$f^*(\Phi) = \sup_{\Lambda \in \mathcal{A}_m} (\langle \Phi, \Lambda \rangle - f(\Lambda)) \quad (76)$$

In view of Theorem 5 in [25], the *biconjugate*  $f^{**}$ , i.e. the conjugate of the conjugate, is equal to the convex hull of  $f$ .

Let  $f(\Lambda) = \text{rank}'(\Lambda)$ . We prove the statement by showing that  $f^{**}$  coincides with (39). The proof consists of two steps.

*Step 1.* Let  $\mathcal{D} := \{\Lambda \in \mathcal{A}_m \text{ s.t. } \|\Lambda\| \leq 1\}$ . Since  $f(\Lambda) = +\infty$  for  $\Lambda \notin \mathcal{D}$ , then its conjugate is

$$\begin{aligned} f^*(\Phi) &= \sup_{\Lambda \in \mathcal{D}} (\langle \Phi, \Lambda \rangle - f(\Lambda)) \\ &= \sup_{\Lambda \in \mathcal{D}} \left( \text{tr} \int \Phi \Lambda^* - f(\Lambda) \right) \end{aligned} \quad (77)$$

where  $\Phi \in \mathcal{A}_m$ . By applying pointwise the *von Neumann's* trace theorem [26], we obtain

$$\int \text{tr}(\Phi \Lambda^*) \leq \int \sum_{k=1}^m \sigma_k(\Phi) \sigma_k(\Lambda) \quad (78)$$

and equality holds if and only if  $\Phi$  and  $\Lambda$  admit the following pointwise *SVDs*:  $\Phi(e^{i\vartheta}) = \Gamma(e^{i\vartheta})\Theta_\Phi(e^{i\vartheta})\Upsilon(e^{i\vartheta})^*$  and  $\Lambda(e^{i\vartheta}) = \Gamma(e^{i\vartheta})\Theta_\Lambda(e^{i\vartheta})\Upsilon(e^{i\vartheta})^*$ . Accordingly,  $f^*$  is independent of  $\Gamma$

and  $\Upsilon$ , therefore

$$f^*(\Phi) = \sup_{\Lambda \in \mathcal{D}} \left( \int \sum_{k=1}^m \sigma_k(\Phi) \sigma_k(\Lambda) - f(\Lambda) \right). \quad (79)$$

If  $\Lambda = 0$ , we have  $f^*(\Phi) = 0$  for each  $\Phi$ . If  $f(\Lambda) = l$ , with  $1 \leq l \leq m$ , then the supremum is achieved by choosing  $\sigma_k(\Lambda(e^{i\vartheta})) = 1$  with  $k = 1 \dots l$ ,  $\vartheta \in [-\pi, \pi]$ , and  $f^*(\Phi) = \int \sum_{k=1}^l \sigma_k(\Phi) - l$ . Thus,  $f^*$  can be expressed as

$$f^*(\Phi) = \int \max \left\{ 0, \sigma_1(\Phi(e^{i\vartheta})) - 1, \dots, \sum_{k=1}^l \sigma_k(\Phi(e^{i\vartheta})) - l, \dots, \sum_{k=1}^m \sigma_k(\Phi(e^{i\vartheta})) - m \right\} \quad (80)$$

and the largest term of this set is the one that sums all positive quantities. We conclude that

$$f^*(\Phi) = \int \sum_{k=1}^r (\sigma_k(\Phi) - 1), \quad (81)$$

where  $r(\vartheta) \in \{0, 1, \dots, m\}$  is such that

$$\begin{cases} r(\vartheta) = 0, & \text{if } \sigma_1(\Phi(e^{i\vartheta})) \leq 1 \\ \sigma_{r(\vartheta)}(\Phi(e^{i\vartheta})) > 1 \text{ and } \sigma_{r(\vartheta)+1}(\Phi(e^{i\vartheta})) \leq 1, & \text{otherwise.} \end{cases} \quad (82)$$

In particular,  $f^*(\Phi) = 0$  for  $\|\Phi\| \leq 1$ .

*Step 2.* We now compute the conjugate of  $f^*$  which is defined as

$$f^{**}(\Lambda) = \sup_{\Phi \in \mathcal{A}_m} (\langle \Lambda, \Phi \rangle - f^*(\Phi)) \quad (83)$$

where  $\Lambda \in \mathcal{A}_m$ . Proceeding as in Step 1, we have

$$f^{**}(\Lambda) = \sup_{\Phi \in \mathcal{A}_m} \left( \int \sum_{k=1}^m \sigma_k(\Lambda) \sigma_k(\Phi) - f^*(\Phi) \right). \quad (84)$$

Next we consider two cases:  $\|\Lambda\| > 1$  and  $\|\Lambda\| \leq 1$ .

• *Case  $\|\Lambda\| > 1$ .* We have,

$$\begin{aligned} f^{**}(\Lambda) &= \sup_{\Phi \in \mathcal{A}_m} \left( \int \left( \sum_{k=1}^m \sigma_k(\Lambda) \sigma_k(\Phi) - \sum_{k=1}^r (\sigma_k(\Phi) - 1) \right) \right) \\ &= \sup_{\Phi \in \mathcal{A}_m} \left( \int \left( \sum_{k=1}^r \sigma_k(\Phi) (\sigma_k(\Lambda) - 1) \right. \right. \\ &\quad \left. \left. + \sum_{k=r+1}^m \sigma_k(\Phi) \sigma_k(\Lambda) + r \right) \right). \end{aligned} \quad (85)$$

Let  $\bar{\vartheta} \in [-\pi, \pi]$  such that  $\|\Lambda\| = \sigma_1(\Lambda(e^{i\bar{\vartheta}})) > 1$ , thus  $\sigma_1(\Lambda(e^{i\bar{\vartheta}})) - 1 > 0$ . Since  $\Lambda \in \mathcal{A}_m$ , then  $\sigma_k(\Lambda(e^{i\vartheta}))$ s are continuous on  $\vartheta \in [-\pi, \pi]$ , thus we can choose  $\sigma_1(\Phi(e^{i\bar{\vartheta}}))$  large enough in a neighborhood of  $\bar{\vartheta}$  so that  $f^{**}(\Lambda) = +\infty$ .

• *Case  $\|\Lambda\| \leq 1$ .* If  $\|\Phi\| \leq 1$ , then  $f^*(\Phi) = 0$  and the supremum is achieved by choosing  $\Phi = I$ , accordingly  $\sigma_k(\Phi(e^{i\vartheta})) = 1$  for each  $\vartheta \in [-\pi, \pi]$ ,  $k = 1 \dots m$ , and

$$f^{**}(\Lambda) = \sum_{k=1}^m \int \sigma_k(\Lambda). \quad (86)$$

Finally, in the case  $\|\Phi\| > 1$  the argument of the sup is always smaller than or equal to the above value:

$$\begin{aligned} & \int \left( \sum_{k=1}^m \sigma_k(\Lambda) \sigma_k(\Phi) - \sum_{k=1}^r (\sigma_k(\Phi) - 1) \right) \\ &= \int \left( \sum_{k=1}^m \sigma_k(\Lambda) \sigma_k(\Phi) - \sum_{k=1}^r (\sigma_k(\Phi) - 1) \right. \\ & \quad \left. - \sum_{k=1}^m \sigma_k(\Lambda) \right) + \sum_{k=1}^m \int \sigma_k(\Lambda) \\ &= \int \left( \sum_{k=1}^m \sigma_k(\Lambda) (\sigma_k(\Phi) - 1) - \sum_{k=1}^r (\sigma_k(\Phi) - 1) \right) \\ & \quad + \sum_{k=1}^m \int \sigma_k(\Lambda) \\ &= \int \left( \sum_{k=1}^r \underbrace{(\sigma_k(\Lambda) - 1)}_{\leq 0 \ \vartheta \in [-\pi, \pi]} \underbrace{(\sigma_k(\Phi) - 1)}_{> 0 \ \vartheta \in [-\pi, \pi]} \right. \\ & \quad \left. + \sum_{k=r+1}^m \sigma_k(\Lambda) \underbrace{(\sigma_k(\Phi) - 1)}_{\leq 0 \ \vartheta \in [-\pi, \pi]} \right) + \sum_{k=1}^m \int \sigma_k(\Lambda) \\ &\leq \sum_{k=1}^m \int \sigma_k(\Lambda) \end{aligned} \quad (87)$$

where we exploited (82).

We conclude that

$$f^{**}(\Lambda) = \begin{cases} \sum_{k=1}^m \int \sigma_k(\Lambda), & \|\Lambda\| \leq 1 \\ +\infty, & \text{otherwise.} \end{cases} \quad (88)$$

■

### C. Proof of Proposition 3.2

Before proving the statement, we establish the following lemma.

*Lemma A.1:* Let  $\mathbf{C}$  be a closed convex subset of  $\{Z \in \mathbf{M}_{m,n} \text{ s.t. } \text{tr}(Z_0) = 0\}$ ,  $c$  be a constant term. If the following convex optimization problem is feasible

$$\begin{aligned} & \max_{\substack{W \in \mathbf{Q}_m \\ Z \in \mathbf{M}_{m,n}}} \log \det W + c \\ & \text{subject to } W \succ 0 \\ & \quad \mathbf{T}(\hat{R}) + \mathbf{T}(Z) \succeq \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix} \\ & \quad Z \in \mathbf{C} \end{aligned} \tag{89}$$

then it admits a solution.

*Proof:* By assumption, the optimization problem is feasible, i.e. there exist  $\bar{W} \in \mathbf{Q}_m$  and  $\bar{Z} \in \mathbf{M}_{m,n}$  satisfying the constraints, and such that  $|\log \det \bar{W} + c| < \infty$ . Accordingly, the above problem is equivalent to maximize  $\log \det W$  over the set

$$\begin{aligned} \mathbf{D} := \{ & (W, Z) \in \mathbf{Q}_m \times \mathbf{C} \text{ s.t. } W \succ 0, \\ & \mathbf{T}(\hat{R}) + \mathbf{T}(Z) \succeq \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix}, \log \det W \geq \log \det \bar{W} \}. \end{aligned}$$

Next we show that  $\mathbf{D}$  is a compact set. Since  $\log \det W$  is continuous over  $\mathbf{D}$ , it follows from *Weierstrass'* theorem that  $\log \det W$  admits a maximum on  $\mathbf{D}$ .

To prove the compactness of  $\mathbf{D}$ , we show that it is bounded and closed. Let  $\{(Z^{(k)}, W^{(k)})\}_{k \in \mathbb{N}}$  be a sequence belonging to  $\mathbf{D}$ . Since the minimum singular value of the map  $\mathbf{T}$  is strictly positive, if  $\|Z^{(k)}\| \rightarrow \infty$  as  $k \rightarrow \infty$ , then  $\|\mathbf{T}(Z^{(k)})\| \rightarrow +\infty$ . Since  $\mathbf{T}(Z^{(k)})$  is a symmetric matrix,  $\mathbf{T}(Z^{(k)})$  has at least one eigenvalue tending to infinity in modulus. Moreover  $\text{tr}(\mathbf{T}(Z^{(k)})) = 0$  because  $Z \in \mathbf{C}$ . Thus  $\mathbf{T}(Z^{(k)})$ , and hence  $\mathbf{T}(\hat{R}) + \mathbf{T}(Z^{(k)})$ , has at least one eigenvalue tending to  $-\infty$ . This is not possible because  $Z^{(k)}$  must satisfy inequality

$$\mathbf{T}(\hat{R}) + \mathbf{T}(Z^{(k)}) \succeq \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix} \succeq 0. \tag{90}$$

Thus,  $\|Z^{(k)}\| < \infty$ . Moreover,  $\|W^{(k)}\| < \infty$  because  $0 \prec W^{(k)} \preceq (\mathbf{T}(\hat{R}) + \mathbf{T}(Z^{(k)}))_{00}$ . Therefore  $\mathbf{D}$  is bounded. Let  $\partial \mathbf{D}$  denote the subset of the boundary of  $\mathbf{D}$  not contained in  $\mathbf{D}$ . Since  $\mathbf{C}$  is a

closed subset of  $\mathbf{M}_{m,n}$ ,  $\partial\mathbf{D}$  is at most the set of elements  $(Z, W)$  such that  $W$  is positive semi-definite and singular. Since  $\lim_{(Z,W) \rightarrow \partial\mathbf{D}} \log \det W = -\infty$  and  $W$  must satisfy the inequality  $\log \det W \geq \log \det \bar{W}$ , we conclude that  $\partial\mathbf{D}$  is an empty set. Accordingly,  $\mathbf{D}$  is closed. ■

We proceed to prove Proposition 3.2. Since  $\mathbf{T}(\hat{R}) \succ 0$ , Problem (49) is feasible (it is sufficient to pick  $W = \alpha I$  and  $Z = 0$  where  $\alpha > 0$  is the minimum eigenvalue of  $\mathbf{T}(\hat{R})$ ). Then, by applying Lemma A.1 with

$$\mathbf{C} := \{Z \in \mathbf{M}_{m,n} \text{ s.t. } \text{diag}(Z_j) = 0 \ j = 0 \dots n, \\ \sum_{j=0}^n |(Z_j)_{kh}| + |(Z_j)_{hk}| \leq \lambda\gamma \ k \neq h, \ \lambda I + \mathbf{T}(Z) \succeq 0\}$$

we conclude that (49) admits a solution. Finally, it is worth noting the objective function in (49) is strictly convex with respect to  $W$ , thus the optimal solution  $W^\circ$  is unique. ■

#### D. Proof of Proposition 3.3

Our proof uses the following lemma whose proof can be found in [11].

*Lemma A.2:* Let  $Z \in \mathbf{M}_{m,n}$ ,  $W \in \mathbf{Q}_m$ . If  $W \succ 0$  and such that

$$\mathbf{T}(Z) \succeq \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix} \quad (91)$$

then  $\mathbf{T}(Z) \succ 0$  and the unique solution to the *Yule-Walker equations*, [27],

$$\begin{cases} \mathbf{T}(Z)B^T = \begin{bmatrix} W \\ 0 \end{bmatrix}, \quad B \in \mathbb{R}^{m \times m(n+1)} \\ B_0 = I \end{cases} \quad (92)$$

is such that  $B\Delta^*$  has zeros inside the unit circle.

We proceed to prove Proposition 3.3. Note that the duality gap between (41) and (49) is equal to zero. In particular,

$$\langle U^\circ, X^\circ \rangle = 0 \quad (93)$$

where  $U^\circ \in \mathbf{Q}_{m(n+1)}$ ,  $U^\circ \succeq 0$  maximizes (48). Note that  $U^\circ$  can be expressed in the following way

$$U^\circ = \mathbf{T}(\hat{R}) + \mathbf{T}(Z^\circ) - \begin{bmatrix} W^\circ & 0 \\ 0 & 0 \end{bmatrix} \quad (94)$$

where  $W^\circ \succ 0$  and  $Z^\circ \in \mathbf{M}_{m,n}$  are solution to Problem (49). By Lemma A.2, we have that  $T(\hat{R}) + T(Z^\circ) \succ 0$ , accordingly  $U^\circ$  has rank at least equal to  $mn$ . Since  $U^\circ, X^\circ \succeq 0$ , (93) implies that  $X^\circ$  has rank at most equal to  $m$ . On the other hand  $\text{rank}(X^\circ) \geq m$  because  $X_{00}^\circ = (W^\circ)^{-1} \succ 0$ . We conclude that  $\text{rank}(X^\circ) = m$ . Hence, there exists  $A \in \mathbb{R}^{m \times m(n+1)}$  full row rank such that  $X^\circ = A^T A$  with  $X_{00}^\circ = A_0^T A_0$ . Since  $U^\circ, X^\circ \succeq 0$ , (93) implies

$$\left( T(\hat{R}) + T(Z^\circ) - \begin{bmatrix} W^\circ & 0 \\ 0 & 0 \end{bmatrix} \right) A^T = 0. \quad (95)$$

By defining  $B \in \mathbb{R}^{m \times m(n+1)}$  such that  $B = A_0^{-1} A$  we obtain

$$(T(\hat{R}) + T(Z^\circ))B^T = \begin{bmatrix} W^\circ \\ 0 \end{bmatrix}, \quad B_0 = I. \quad (96)$$

Since  $T(\hat{R}) + T(Z^\circ) \succ 0$ , the Yule-Walker equations (96) admits a unique solution such that  $B\Delta^*$  has zeros inside the unit circle. Accordingly, there exists  $X^\circ$  such that

$$\Delta X^\circ \Delta^* = \Delta A^T A \Delta^* = (\Delta B^T)(W^\circ)^{-1}(B\Delta^*) \succ 0. \quad (97)$$

Finally, uniqueness of  $X^\circ$  follows from the uniqueness of  $W^\circ$  and  $B$ . It remains to be shown the existence of  $L^\circ$ . In view of (41), we have

$$\begin{aligned} L^\circ &= \arg \min_{L \in \mathbf{Q}_{m(n+1)}} \lambda \gamma h_\infty(D(X^\circ + L)) + \lambda \text{tr}(L) \\ &\text{subject to } L \succeq 0 \end{aligned} \quad (98)$$

where the objective function is continuous. Since  $L = 0$  is a feasible point, we can restrict  $L$  to belong to

$$\begin{aligned} \mathbf{D} &:= \{L \in \mathbf{Q}_{m(n+1)} \text{ s.t. } L \succeq 0, \\ &\lambda \gamma h_\infty(D(X^\circ + L)) + \lambda \text{tr}(L) \leq \lambda \gamma h_\infty(D(X^\circ))\}. \end{aligned} \quad (99)$$

It is not difficult to show that  $\mathbf{D}$  is a closed and bounded set, therefore by *Weierstrass'* theorem  $L^\circ$  does exist. ■

### E. Proof of Proposition 3.4

By Proposition 3.3,  $X^\circ$  is unique. If (56) admits a unique solution  $H$ , then  $L^\circ = G^T H G$  is unique. Therefore,  $\mathcal{V}_{E_m}$  and  $\mathcal{V}_G$  are unique because the uniqueness of  $X^\circ$  and  $L^\circ$ . Equation (56) may be written in the compact form

$$Ay = b \quad (100)$$

where the vector  $y \in \mathbb{R}^{l(l+1)/2}$  contains the independent parameters of  $H$ ,  $A \in \mathbb{R}^{(n+1)|I| \times l(l+1)/2}$  only depends on  $G$  and  $b \in \mathbb{R}^{(n+1)|I|}$  only depends on  $X^\circ$ . If (56) admits a unique solution, then it is obtained in the following way

$$y = (A^T A)^{-1} A^T b \quad (101)$$

and changing  $b$  (i.e.  $X^\circ$ ) such a solution is still unique. Accordingly, the uniqueness of the solution to (56) is equivalent to the uniqueness of the decomposition

$$\Phi_m^{-1} = \Sigma - \Lambda \quad (102)$$

with  $\Phi_m^{-1} \in \mathcal{V}_{E_m} + \mathcal{V}_G$ ,  $\Sigma \in \mathcal{V}_{E_m}$  and  $\Lambda \in \mathcal{V}_G$ . Therefore,  $\mathcal{V}_{E_m} \cap \mathcal{V}_G = \{0\}$ . ■

### F. Proof of Proposition 4.1

The statement can be proved by duality theory along the same line of the proof of Proposition 3.2 and Proposition 3.3, respectively. ■

## REFERENCES

- [1] S. Lauritzen, *Graphical Models*. Oxford: Oxford University Press, 1996.
- [2] A. Willsky, "Multiresolution markov models for signal and image processing," in *Proc. IEEE*, vol. 90, no. 8, Aug. 2002, pp. 1396–1458.
- [3] V. Chandrasekaran, P. Parrilo, and A. Willsky, "Latent variable graphical model selection via convex optimization," *Annals of Statistics (with discussion)*, vol. 40, no. 4, pp. 1935–2013, Apr. 2010.
- [4] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM J. Optim.*, vol. 21, no. 2, pp. 572–596, Feb. 2011.
- [5] R. Otazo, D. Sodickson, and E. Candès, "Low-rank + Sparse (L+S) reconstruction for accelerated dynamic MRI with separation of background and dynamic components," *Proc. SPIE*, vol. 8858, pp. 88 581Z1–88 581Z8, 2013.
- [6] M. Choi, V. Chandrasekaran, and A. Willsky, "Gaussian multiresolution models: exploiting sparse markov and covariance structure," *IEEE Trans. Signal Processing*, vol. 58, no. 3, pp. 1012–1024, Mar. 2010.
- [7] J. Saunderson, V. Chandrasekaran, P. Parrilo, and A. Willsky, "Tree-structured statistical modeling via convex optimization," in *Proc. 50th IEEE Conference on Decision and Control*, Dec. 2011, pp. 2883–2888.



- [8] J. Burg, *Maximum entropy spectral analysis*. PhD Thesis, Stanford Univ., 1975.
- [9] C. Byrnes, T. Georgiou, and A. Lindquist, “A new approach to spectral estimation: A tunable high-resolution spectral estimator,” *IEEE Trans. Signal Processing*, vol. 48, no. 11, pp. 3189–3205, Nov. 2000.
- [10] J. Songsiri and L. Vandenberghe, “Topology selection in graphical models of autoregressive processes,” *J. Mach. Learning Res.*, vol. 11, pp. 2671–2705, 2010.
- [11] J. Songsiri, J. Dahl, and L. Vandenberghe, “Graphical models of autoregressive processes,” in *Convex Optimization in Signal Processing and Communications*, D. Palomar and Y. Eldar, Eds. Cambridge: Cambridge Univ. Press, 2010, pp. 1–29.
- [12] E. Avventi, A. Lindquist, and B. Wahlberg, “ARMA identification of graphical models,” *IEEE Trans. Autom. Control*, vol. 58, no. 5, pp. 1167–1178, May 2013.
- [13] T. Cover and J. Thomas, *Information Theory*. New York: Wiley, 1991.
- [14] P. Stoica and R. Moses, *Introduction to spectral analysis*. New Jersey: Prentice Hall, 1997.
- [15] T. Kailath, A. Sayed, and B. Hassibi, *Linear estimation*. Prentice Hall, 2000.
- [16] C. Byrnes, S. Gusev, and A. Lindquist, “A convex optimization approach to the rational covariance extension problem,” *SIAM J. Optim.*, vol. 37, no. 1, pp. 211–229, 1998.
- [17] —, “From finite covariance windows to modeling filters: A convex optimization approach,” *SIAM Rev.*, vol. 43, pp. 645–675, 2001.
- [18] A. Lindquist and G. Picci, “Linear stochastic systems: A geometric approach to modeling, estimation and identification,” To appear in 2015. Preprint available in <http://www.math.kth.se/optsys/forskning/forskarutbildning/5B5715/LPbook.pdf>.
- [19] R. Dahlhaus, “Graphical interaction models for multivariate time series,” *Metrika*, vol. 51, no. 2, pp. 157–172, Feb. 2000.
- [20] D. Brillinger, “Remarks concerning graphical models for times series and point processes,” *Revista de Econometrica*, vol. 16, pp. 1–23, 1996.
- [21] M. Deistler and C. Zinner, “Modelling high-dimensional time series by generalized linear dynamic factor models: An introductory survey,” *Communications in Information & Systems*, vol. 7, no. 2, pp. 153–166, 2007.
- [22] L. Ahlfors, *Complex Analysis*. New York: McGraw-Hill, 1953.
- [23] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge: Cambridge Univ. Press, 2004.
- [24] A. Abdelwahab, O. Amor, and T. Abdelwahed, “The analysis of the interdependence structure in international financial markets by graphical models,” *Int. Res. J. Finance Econ.*, vol. 15, pp. 291–306, 2008.
- [25] R. Rockafellar, *Conjugate duality and optimization*. No. 16 in Conference Board of Math. Sciences Series, 1974.
- [26] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge: Cambridge Univ. Press, 1990.
- [27] T. Söderström and P. Stoica, *System Identification*. UK: Prentice-Hall, 1989.